Oracle生成式AI和检索增强生成解决方案

公益讲座11：00准时开始，请大家先浏览云技术微信公众号技术文章。资料会在各群同步发布，已入群客户请勿重复入群！

20-23

数据库和云讲座群

甲骨文云技术公众号

B站专家系列课程

基于 Oracle 数据库
免费企业数据健康检查

- 及时了解数据库健康状况，发现并解决潜在问题
- 维护数据库系统良好状态，保护数据资产的安全
- 提升数据库性能、稳定性和安全性，降低业务风险

免费咨询热线：

400-699-8888

*活动最终解释权归甲骨文公司所有

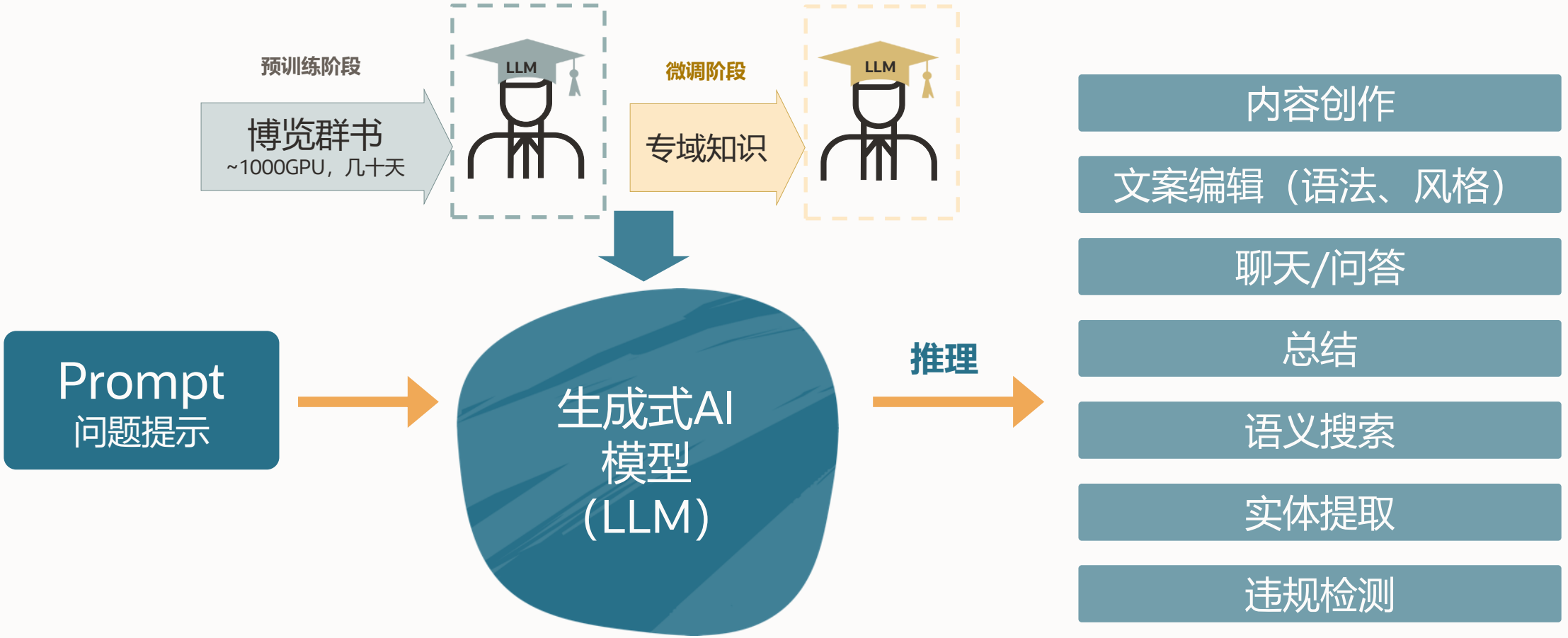
Oracle生成式AI和检索增强生成解决方案

Oracle China
刘群策

议程

- Oracle生成式AI和相关服务介绍
- Oracle向量数据库和检索增强生成方案
- Oracle加速企业生成式AI落地和创新
- 讨论

当前热点话题和技术：生成式AI和大语言模型LLM



生成式AI（Generative AI），又称为**AIGC（AI生成内容），**指通过机器学习、深度学习或其他相关的AI技术，使用算法和模型来生成各种形式的内容，如图像、音频、视频、文本等。AI文章概括总结和续写，文字转图像，AI数字化主持人等，都属于生成式AI的范畴。

大语言模型（LLM）指的是生成式AI中的参数巨大的自然语言模型。

Oracle生成式AI战略

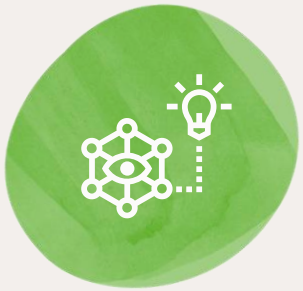
目前很少有通过生成式AI来满足企业客户的端到端需求

Oracle通过生成式AI来转化业务体验

Oracle 应用

嵌入在甲骨文云应用和数据库产品组合中的 AI

Oracle将在Oracle的云应用、行业应用，其他AI服务和数据库产品组合中嵌入生成式AI能力。

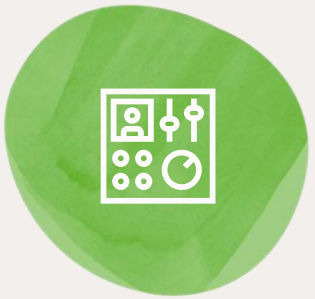


OCI 生成式AI云服务

可微调的LLM和支持检索增强的AI服务

OCI客户可直接使用生成式AI能力，支持微调和部署自定义模型。

提供RAG框架相关服务和加速开发包，以及向量检索的数据库能力



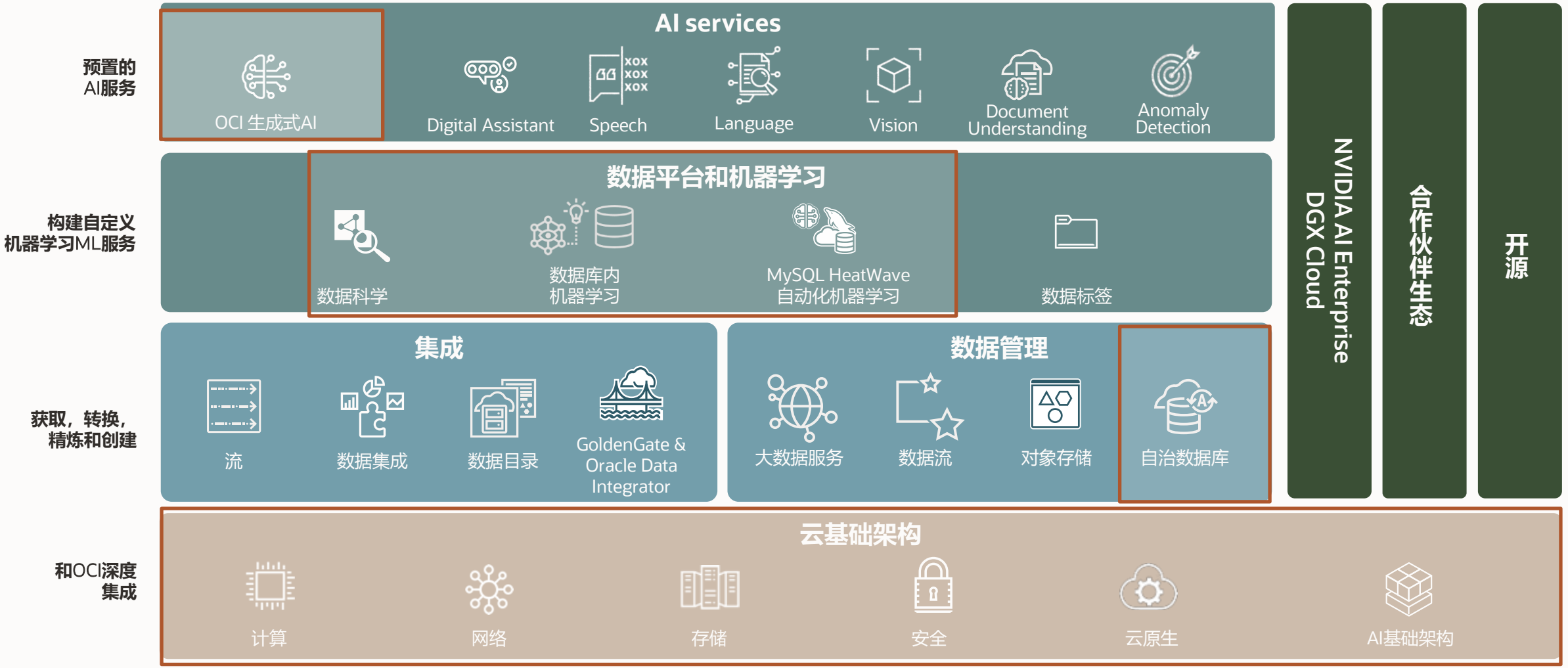
GPU 基础架构

用于模型训练的高速、低成本 AI 基础设施

LLM，包括Cohere创建的LLM，需要在OCI上建立大规模的AI基础设施，以经济高效的方式训练，微调和部署其生成式AI模型。

Oracle 完整的企业级 AI 能力

AI 应用 – 自定义, Fusion, Cerner, NetSuite, 垂直应用



Oracle & Nvidia 助力大模型训练和机器学习

OCI基础架构优势：

- 各种GPU资源，满足不同需求
- 高速网络
- 成本
- 数据管理和安全
- 管理维护

Microsoft　mosaicML　AON　RAZER

Qti.ai　Adept　SoundHound　EMORY UNIVERSITY

大语言模型的训练，部署，推理

OCI 生成式AI / AI Service	OCI数据科学服务	Nvidia AI Enterprise Nvidia DGX Cloud

OCI（GPU & Supercluster）

OCI更好地支持大模型训练和推理

- 提供超级集群 (OCI Superclusters)，提供基于融合以太网 (RoCE) v2上的RDMA 的超级集群，15微秒延迟，1600Gbps。
- 目前支持单集群最大4096节点（32,768 个NVIDIA A100 GPU）
- NVIDIA H100 GPU：新一代GPU将在OCI提供(节点间3200Gbps，支持更多节点)，并提供NVIDIA AI Enterprise，包括AI工作流程每个步骤的基本处理引擎，从数据处理和AI模型训练到模拟和大规模部署。

VM GPU.A10.1	VM GPU.A10.2	BM GPU.A10.4	BM GPU4.8	BM GPU.A100-v2.8
A10	A10	A10	A100	A100
1	2	4	8	8
24 GB	48 GB	96 GB	320 GB	640 GB
			1.6 Tb/sec RDMA	1.6 Tb/sec RDMA
$2.00 per GPU/hr	$2.00 per GPU/hr	$2.00 per GPU/hr	$3.05 per GPU/hr	$4.00 per GPU/hr

Oracle生成式AI服务及优势

ORACLE
+
cohere

Command模型	Summarize模型	Embed模型
•此模型可接收用户的提示并生成文本。Command 有两种不同的大小，可根据业务用例进行高度定制化，包括文本生成、文本汇总、RAG 和聊天。	•此模型可对文本进行抽象汇总，支持用户根据特殊用例需求，使用各种参数来配置结果。	•可将文本转换为数值向量。提供了英语和多语言模型（支持 100 多种语言），包括语义搜索、文本分类、RAG 搜索引擎和旧版搜索改进。

Oracle生成式AI Agent RAG Service*

端到端，简化为企业设计

可针对客户数据定制

安全和隐私

API调用或独立部署

OCI AI Services概览
结合生成式AI，支持更多创新

Language 语言
- 语言检测
- 实体识别
- 类别识别
- 关键短语提取
- 情感分析

Speech 语音
- 语音转文本
- 语音翻译
- 文本转语音
- 识别发言者
- 发言者语气情绪识别

Vision 图像
- 图像说明
- 图像分类
- 物体识别（可能性和坐标）
- 人脸识别（部位坐标）

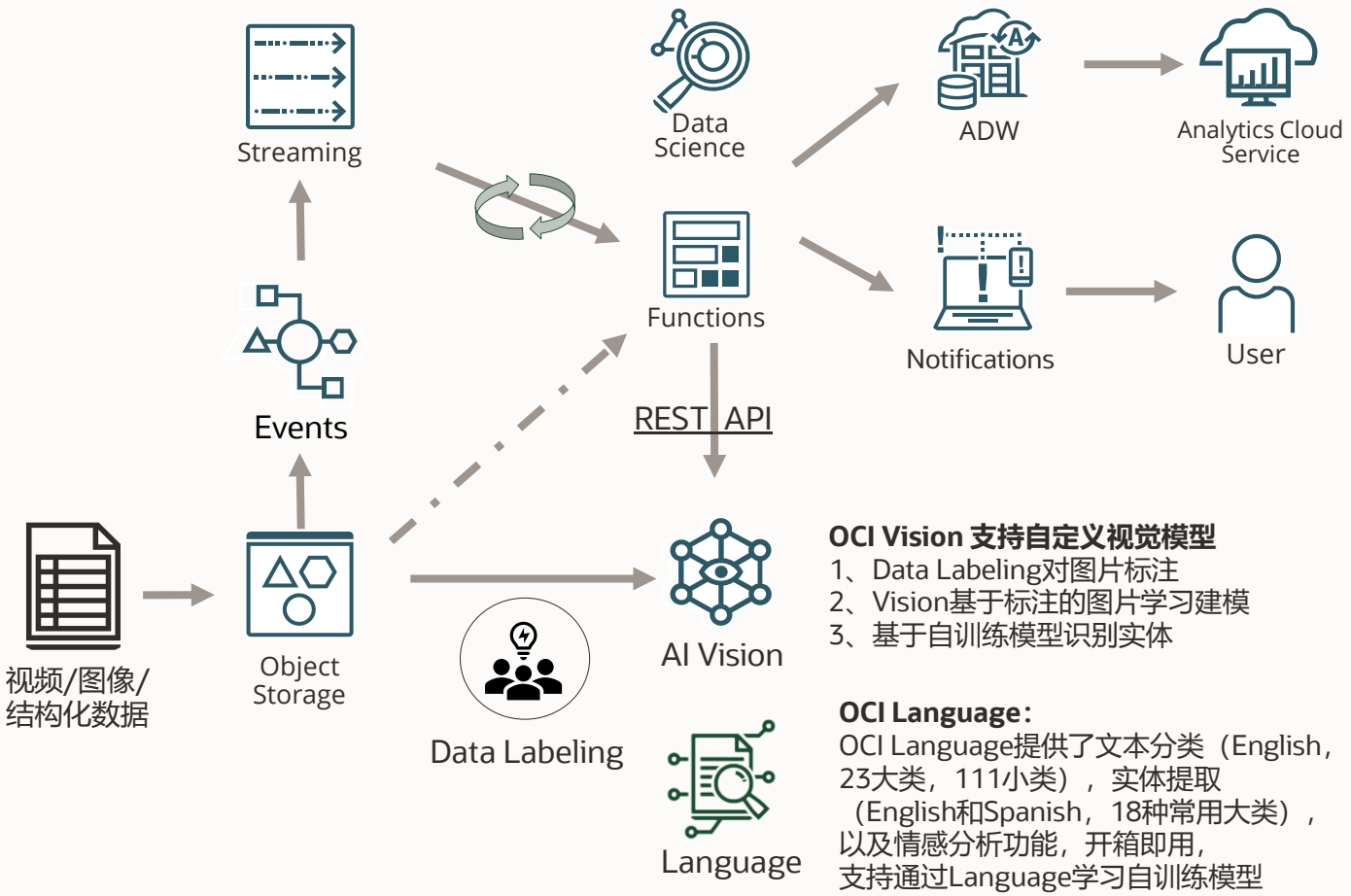
Anomaly Detection
- 分析大量相关数据并以最大的准确性尽早识别出异常
- 时间序列数据
- 提供单变量和多元内核

Document Understanding
- OCR-图像转文本/标签
- 图片文本提取
- 图像转数据(K-V, 表格, 地址等...)
- 图片根据文档分类
- OCR-PDF转文本

Data Labeling 数据标签
- 创建和浏览数据集，查看数据记录（文档、文本和图像）
- 标识文档、文本和图像（记录）的属性（标签）

预置模型的AI服务，可扩展，使用方式：Console, REST APIs, SDK, CLI

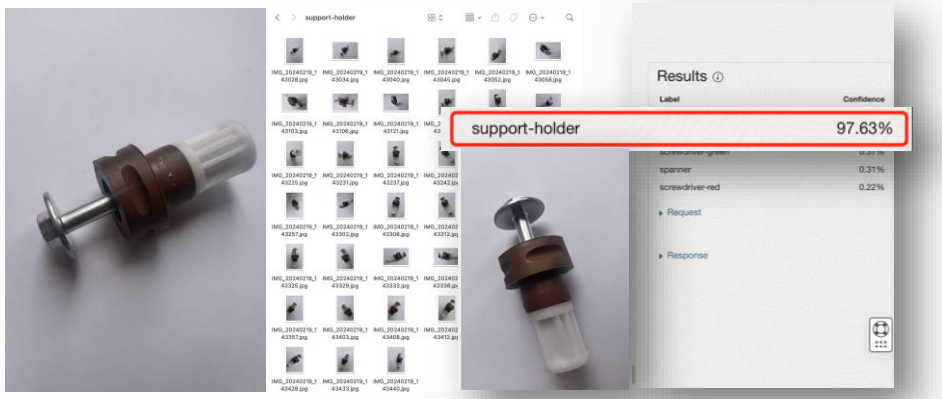
基于OCI AI Service构建自己的智能应用



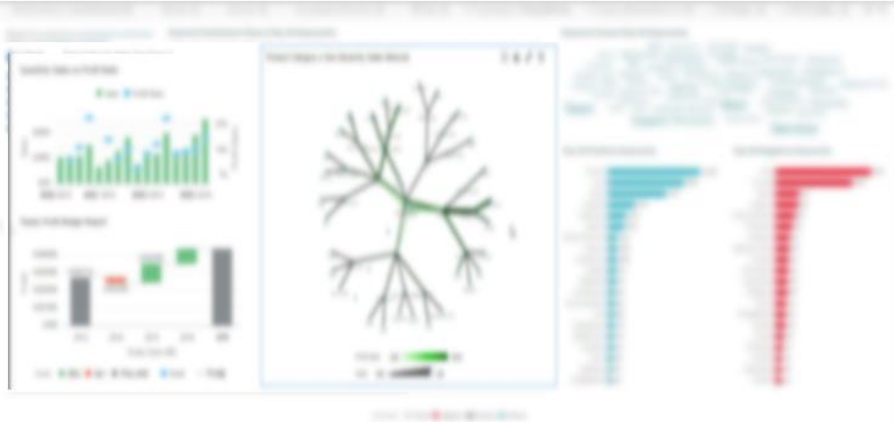
OCI Vision 支持自定义视觉模型
1、Data Labeling对图片标注
2、Vision基于标注的图片学习建模
3、基于自训练模型识别实体

OCI Language：
OCI Language提供了文本分类（English，23大类，111小类），实体提取（English和Spanish，18种常用大类），以及情感分析功能，开箱即用，
支持通过Language学习自训练模型

识别机器配件或工具



针对客户评价和反馈进行分类，情感分析
自动化处理客户反馈，提高业务效率和客户满意度。



议程

- Oracle生成式AI和相关服务介绍
- Oracle检索增强生成方案和示例
- Oracle加速企业生成式AI落地和创新
- 讨论

生成式AI和大语言模型面临的挑战

幻觉

听起来似是而非的虚假信息

自信的响应无法通过训练数据证明

安全

模型操纵风险：深度伪造和恶意攻击

数据安全风险：隐私泄露或知识产权窃取

实时性

基于过去的历史数据学习和推理

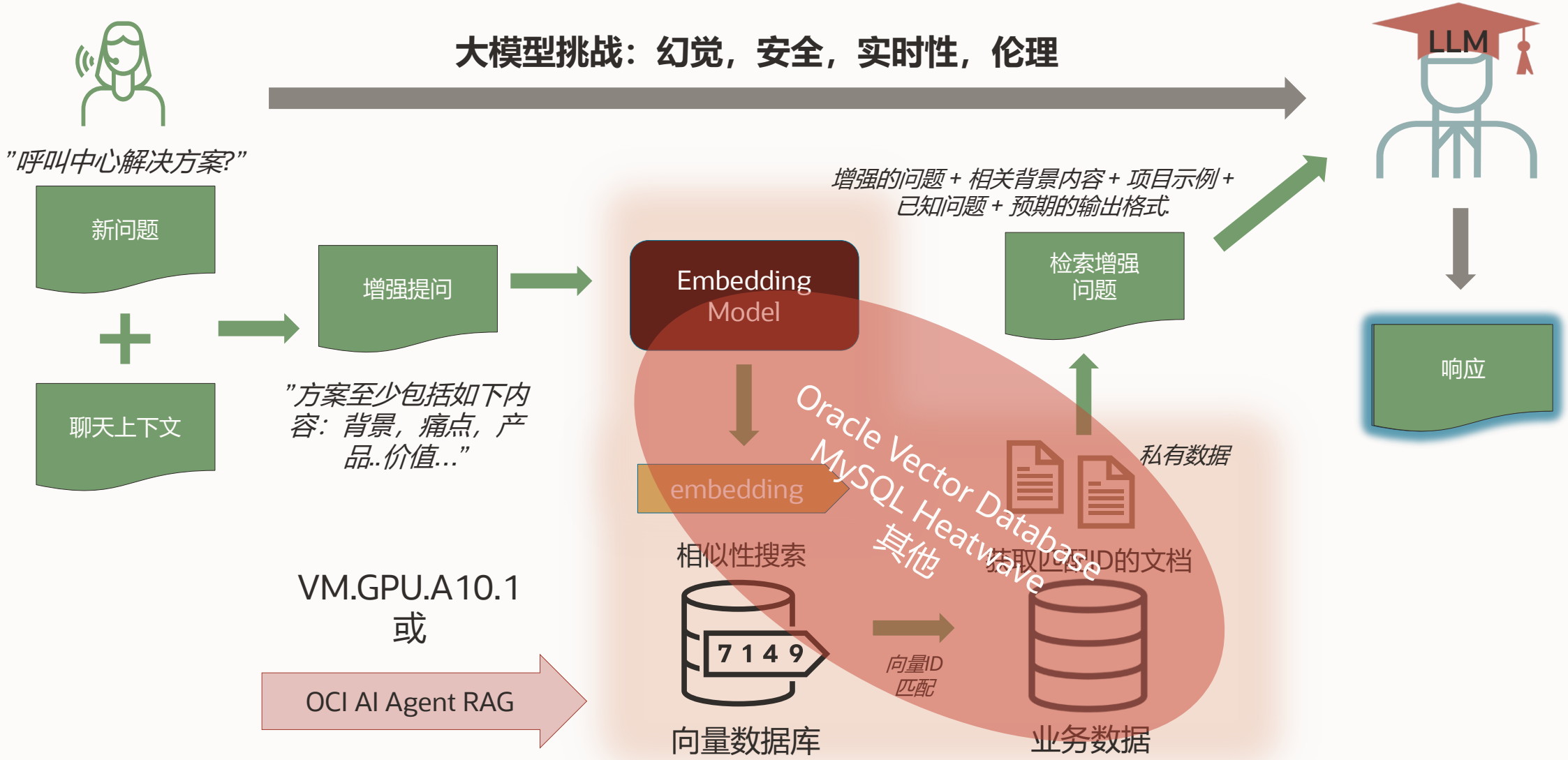
不了解私有信息或企业内部信息

伦理

偏见和问责

透明度和社会影响

利用检索增强生成(RAG)来提升生成结果的准确性和实时性



为什么检索增强使用向量更加准确？？

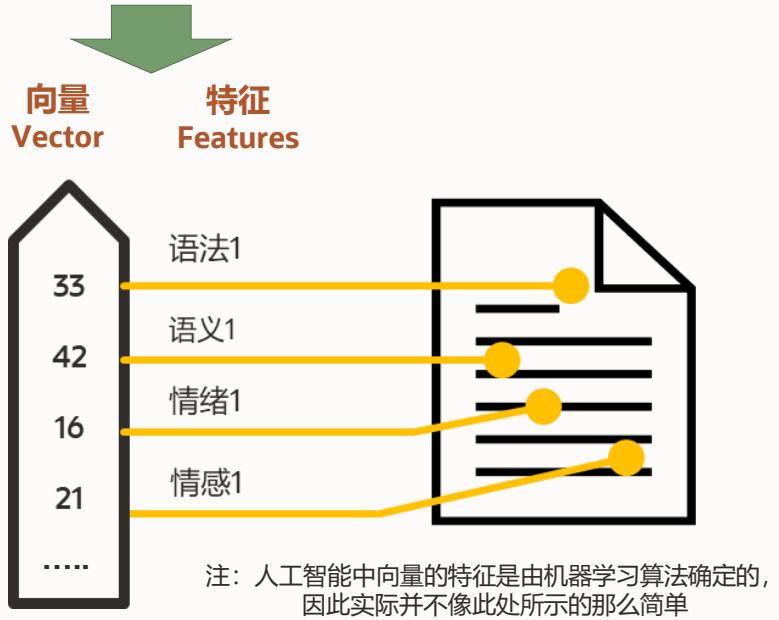
向量（Vector）

- 向量（也称为欧几里得向量、几何向量），指具有大小和方向的量。可以使用带箭头的线段表示，箭头指向即为向量的方向，线段的长度表示向量的大小。两个向量的距离或者相似性可以通过汉明距离、欧式距离或者余弦距离得到。

向量嵌入（Vector Embedding）

- 一种自动化提取事物特征值的方法，用来生成高维度的向量数据
- 图像、文本和音视频这种非结构化数据都可以通过某种变换或者嵌入深度学习转化为向量数据。
- 使用深度学习嵌入模型来生成向量数据。例如，文本向量可以通过 Oracle生成式AI的 Embedding模型生成。

机器抽象出成百上千个维度（数字）代表文档的不同特征



注：人工智能中向量的特征是由机器学习算法确定的，因此实际并不像此处所示的那么简单

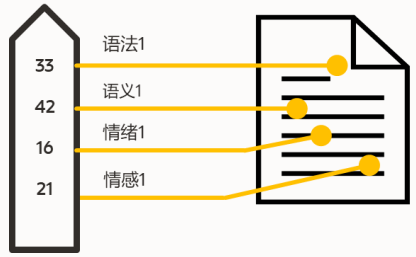
例如：我们将这句话 “好的企业级数据库解决方案” 用 模型进行文本 Embedding，它会生成一个 >1000 维的向量，得到的结果是这样：“[0.011199951,0.038208008, … …, 0.007411957,-0.027008057]” 。嵌入通常存储在向量数据库中。其搜索功能侧重于文本的含义，而不是仅仅根据关键字查找结果。向量检索用于语义搜索，文本分类，文本聚类和推荐系统。

Oracle 数据库 23c：为检索增强生成（RAG）提供企业级支撑

向量数据库和AI向量搜索助力构建企业级知识库

- 原生支持向量数据的存储和高效搜索
- 新的SQL Embedding函数用于生成向量数据
- 新的VECTOR 数据类型用于存储向量数据
- 新的 SQL 语法和函数用于向量相似性搜索
- 新的近似搜索索引经过打包和调优以实现高性能和高质量搜索
- 在查询中与相关客户和产品的业务数据一起执行向量搜索
- 在同一数据库中处理向量和其他工作负载

例如“好的数据库企业级解决方案”，embedding以后，生成一个>1000维的向量，其搜索功能侧重于文本的含义，而不是仅仅根据关键字查找结果。




```
CREATE TABLE tbl_kms_solution (km_id    number,
                               Industry varchar2(400),
                               catalog varchar2(400),
                               doc blob,
                               text_vec vector,
                               ……
);
```

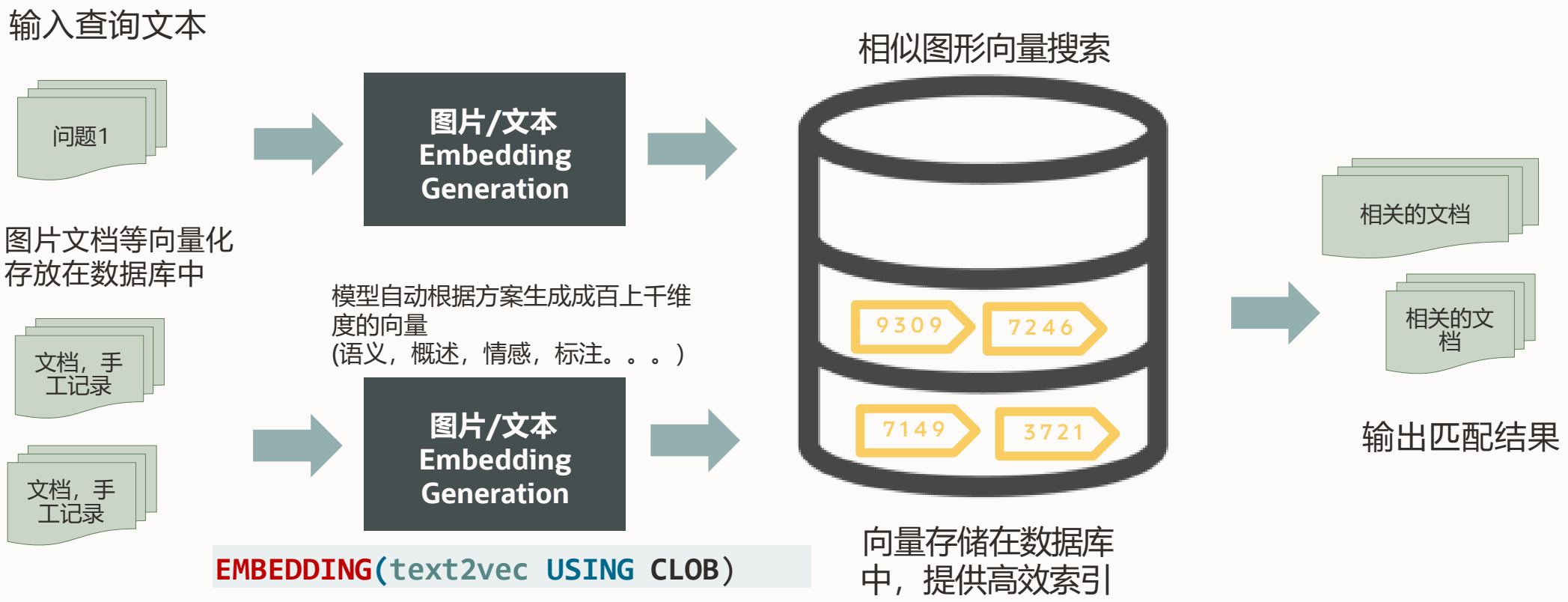
将客户数据、方案数据和 AI 搜索结合在 5 行 SQL 代码中！

```
SELECT …
FROM   tbl_kms_solution
WHERE  investment >= (SELECT budget FROM customer …)
AND    industry   in (SELECT  industry FROM customer …)
ORDER BY vector_distance(text_vec, :input_vector);
```

Oracle 数据库 23c：向量数据库和AI向量搜索构建企业知识库

原生支持生成向量数据，提供 SQL EMBEDDINGS，相似性搜索等函数

```
select id, doc from tbl_kms order by VECTOR_DISTANCE(text_vec,
EMBEDDING(text2vec USING :input_text)) fetch first 2 rows only;
```



Oracle向量数据库和AI向量搜索的独特优势

结合高质量业务数据，支持复杂的融合 SQL和高价值结果

Oracle 融合数据库，支持所有类型的工作负载和数据模型如图、文本、JSON、地理信息、关系型等，也支持所有 SQL，包括复杂的运算和功能

只有Oracle 数据库才能把向量数据搜索和关系型数据一起合成复杂的、融合的SQL查询，产生业务价值

找到符合条件的项目资料（立项/案例……），需要和当前项目或产品有关系，过去五年内，按照产品和供应商分组统计项目个数，至少有超过5个成功实施案例。
排名前3个相似的项目材料
还可以有更多业务条件，例如文档权限，部门，行业，类型……

专用向量数据库很难做到这一点，往往要多次查询和结合其他

Top-3
(top 3 document per matching group)

Vector Search
(和询问的问题相似的内容)

Having Clause
Having count > 5

Group by
(按供应商分组, 统计数量)

Graph
(有关联关系)

Relational
(过去5年)

Oracle AI Vector 存储和检索能力扩展

Oracle数据库所有核心功能都可使用，实现企业级性能，可靠性和敏捷性

RAC 集群

集群，高可用，扩展

APEX

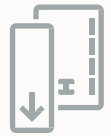
低代码开发，应用

分区

数据分布，加速查询

并行执行

加速查询和处理



事务

数据一致性

分析

模式匹配，ML，多维

安全

加密，脱敏，审计

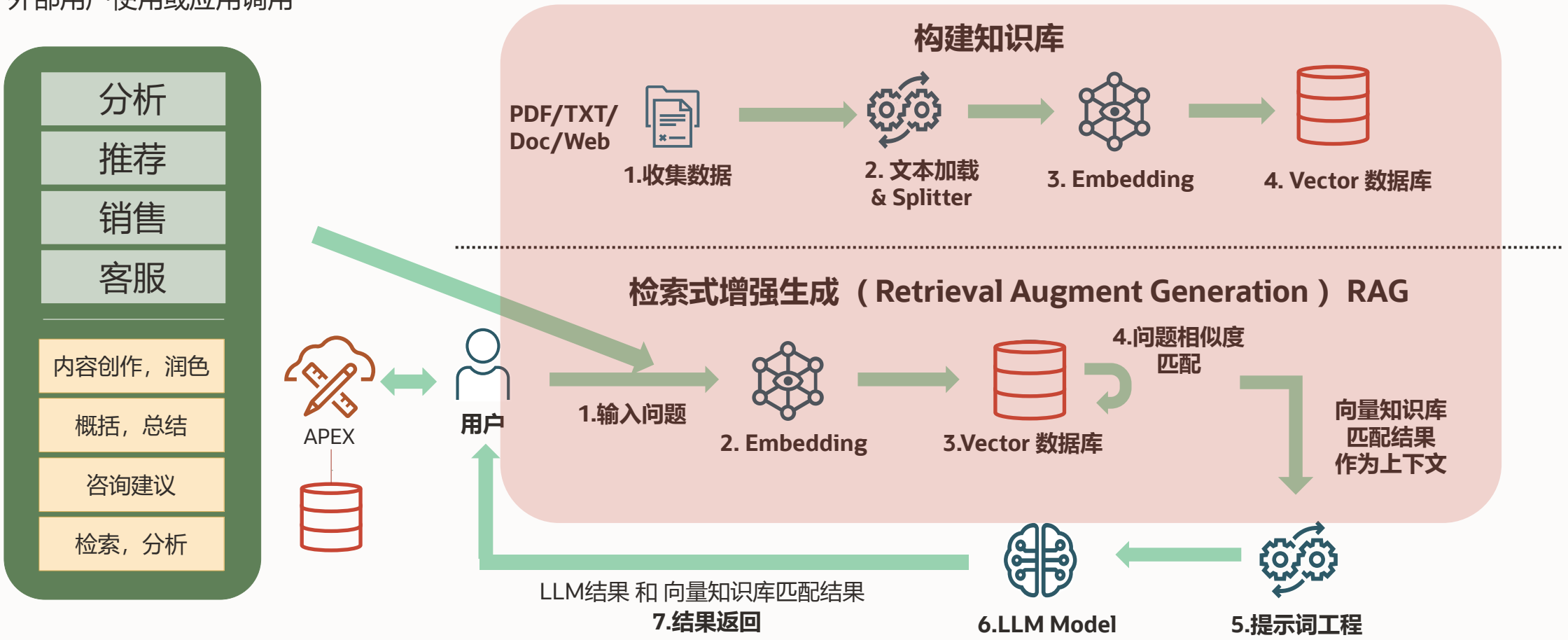
Exadata

预置加速优化，整合

Oracle向量数据库能力，支持企业实现增强检索生成

提升生成式AI的准确性，实时性和与企业/行业的结合度，加速与企业其他应用集成

如产品说明文档、FAQ、客服历史记录等材料，借助LLM + Embedding + Vector Database等技术构建企业专属知识库，提供给内部员工，外部用户使用或应用调用



Oracle数据库的Select AI

自然语言自动转化为SQL，在数据库中执行查询

设置上下文

识别源表和元数据

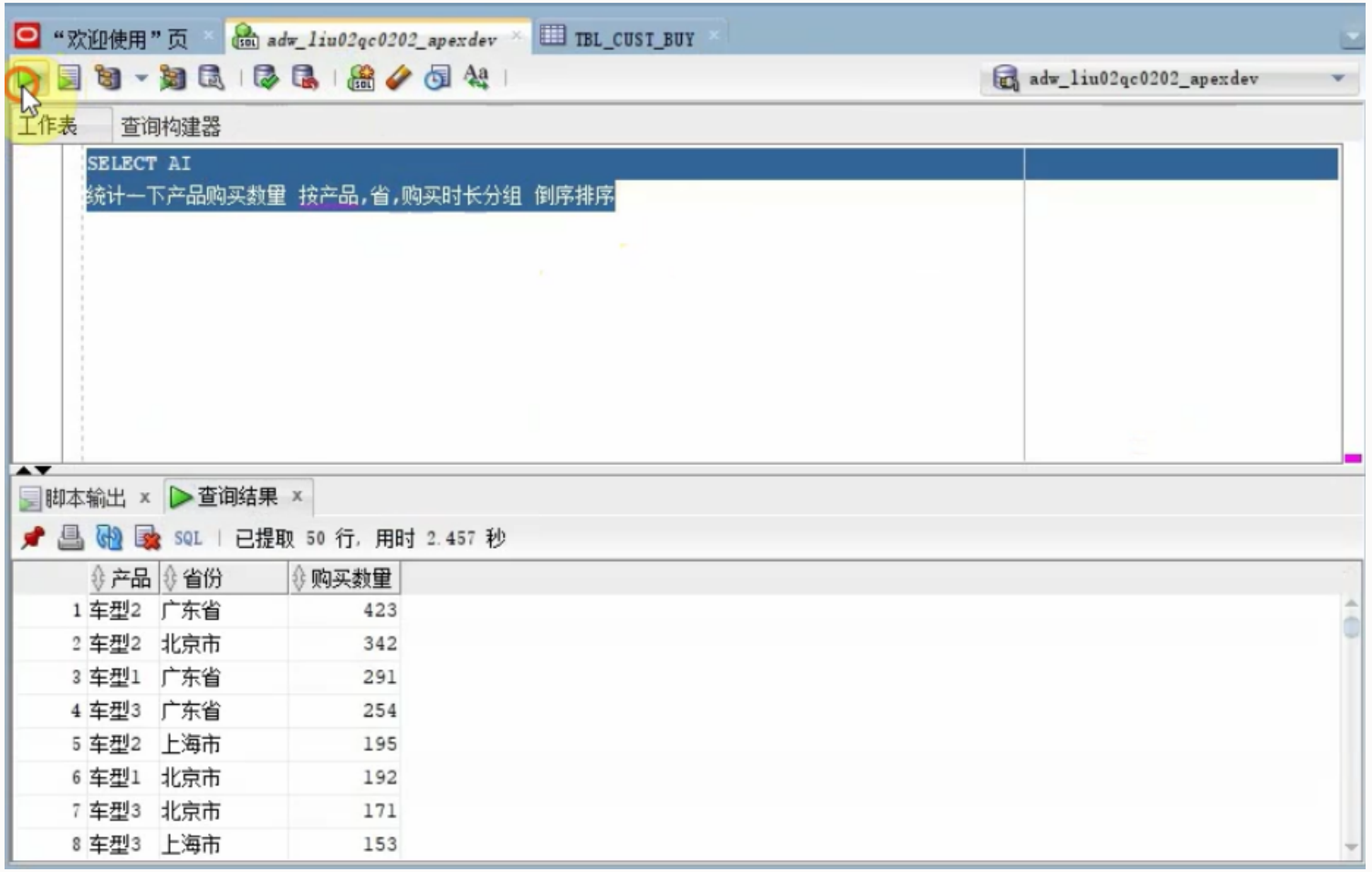
构建提示

提示=说明+上下文+
用户问题

发送提示
以生成 SQL

返回SQL

执行



零售行业
分析型/生成式AI案例

新产品

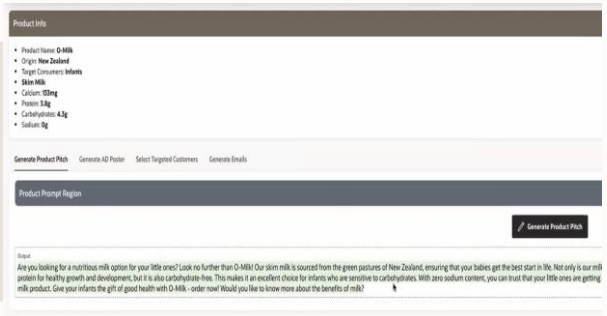
Marketing Mgr.

Customer

1

- 输入产品属性
- **生成产品介绍**
- **生成产品广告词**

【OCI Generative AI 】



2

生成产品广告图片

【Stable Diffusion】

3

基于数据库内数据分析

- 找到类似产品中销售最好的产品
- **利用机器学习分群和预测找到目标客户**【select AI】
- **自动生成邮件草稿** 基于产品属性和客户信息【OCI Generative AI】
- 发送推荐邮件给目标客户

4

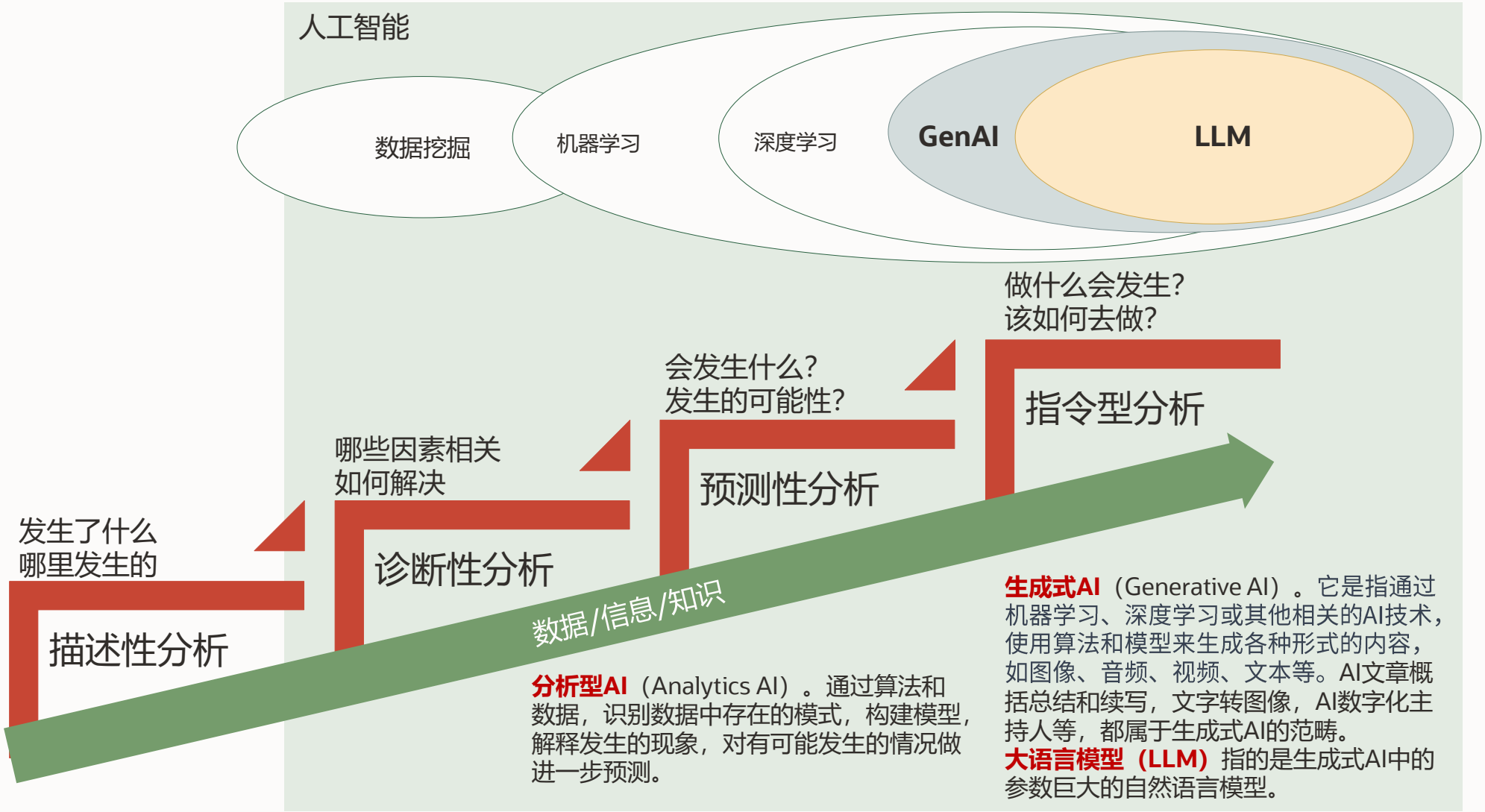
客户 Q & A 通过 RetailGPT

【 OCI Generative AI 】

议程

- Oracle生成式AI和相关服务介绍
- Oracle检索增强生成方案和示例
- Oracle加速企业生成式AI落地和创新
- 讨论

数据分析和人工智能在企业的应用



企业生成式AI架构与部署方式

从生成式AI在企业端到端的应用角度，需要考虑的能力和要求

如何引入生成式AI?

- 语言和模态
- 微调
- 预训练
- 部署方式

如何用好生成式AI?

- AI Agent
- 检索增强生成
- 知识库
- 数据

如何与生成式AI交互

- 聊天机器人
- 工具
- 流程
- 开发

企业生成式AI架构与部署方式??

从生成式AI在企业端到端的应用角度，需要考虑的能力和要求

如何引入生成式AI?

- 语言和模态
- 微调
- 预训练
- 部署方式

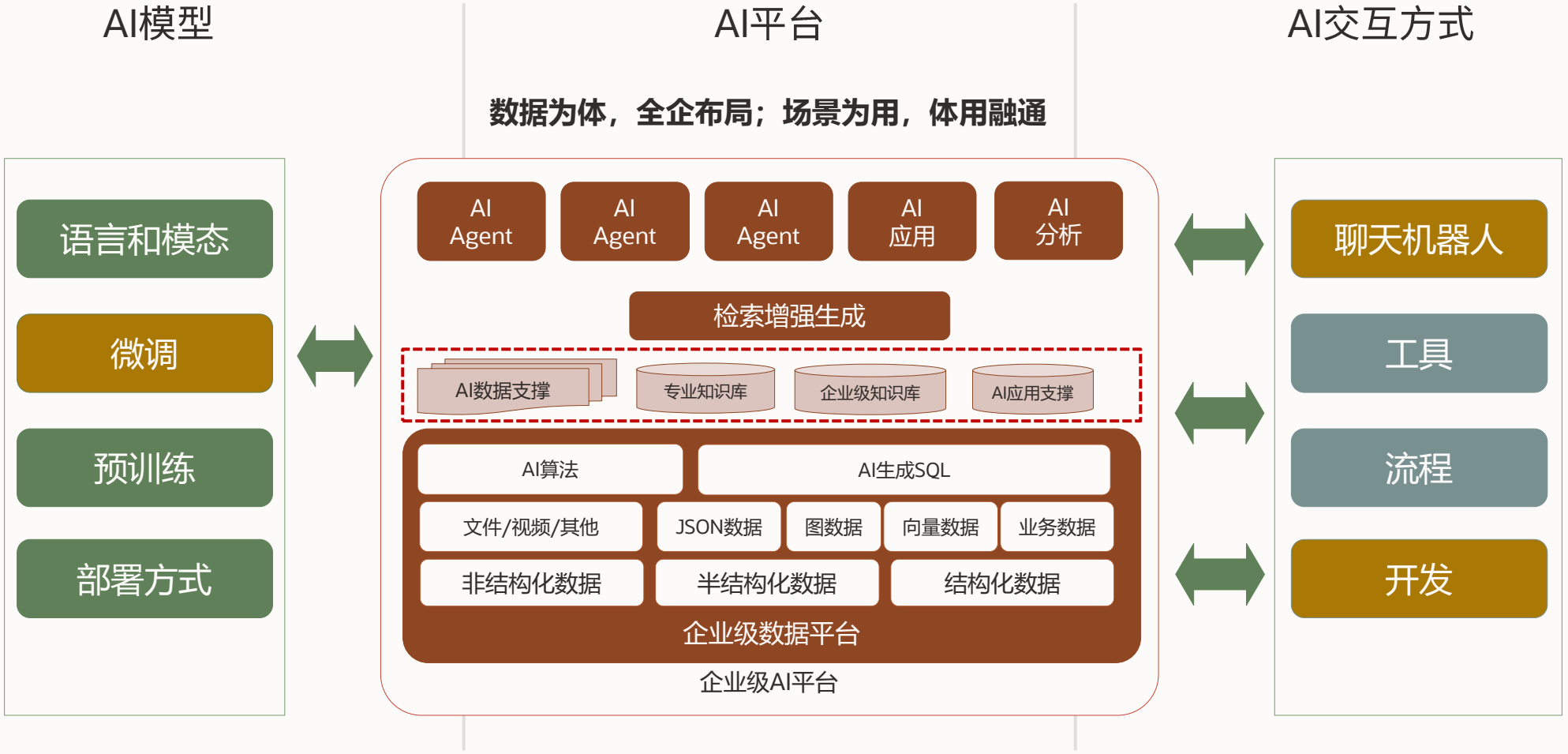
如何用好生成式AI?

如何与生成式AI交互?

- 聊天机器人
- 工具
- 流程
- 开发

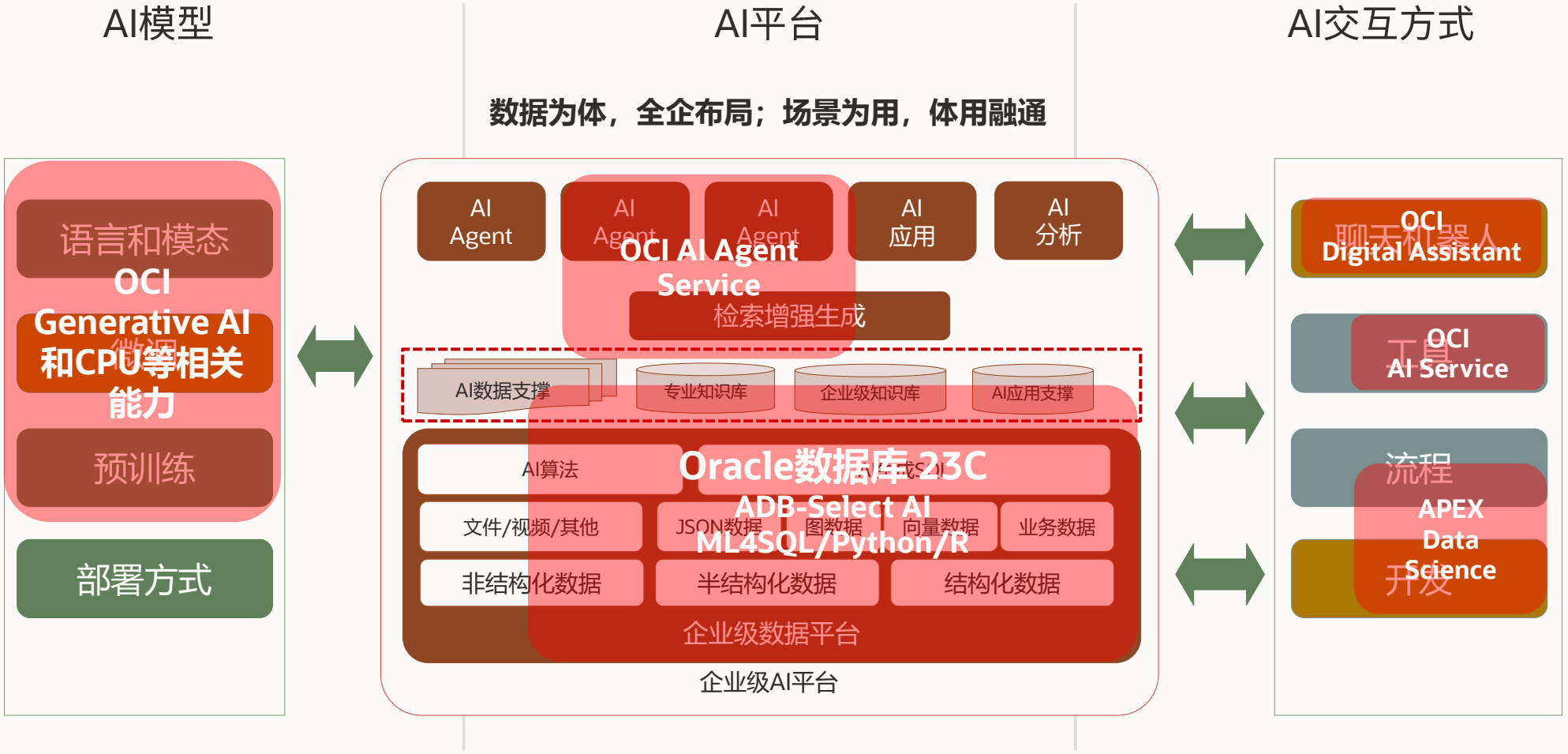
企业生成式AI架构与部署方式

从生成式AI在企业端到端的应用角度，需要考虑的能力和要求

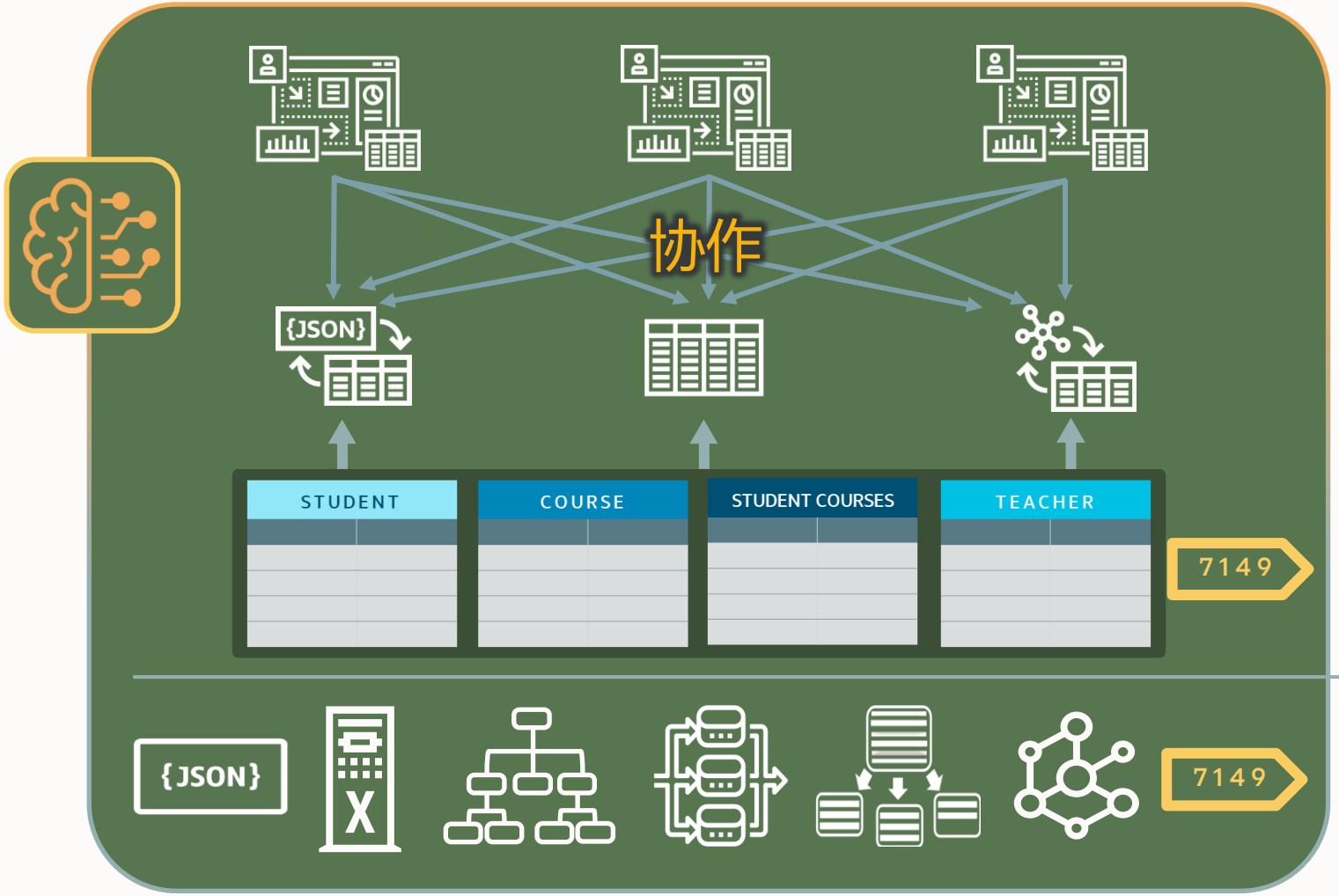


企业生成式AI架构Oracle解决方案

从生成式AI在企业端到端的应用角度，需要考虑的能力和要求



Oracle数据库 23C：让生成和运行任何规模的应用都很简单



生成式人工智能和RAG

允许自然语言提出问题和意图，自动生成 数据访问模型（JSON 二元性）、 低代码开发（APEX 蓝图）、数据查询（SQL） 等。

为生成式AI提供检索增强生成（RAG）支持，结合企业自有数据，提升大语言模型的准确性和信息实时性

为大语言模型提供各种格式的高质量数据

结构化，JSON，地理，属性图，向量

使用融合数据库引擎进行丰富的基于SQL的分析

仅仅是添加 SQL 语句(或REST API调用），而不是另一个数据库

直接对JSON存储和联合查询商圈购买力和销售情况

```
SELECT SUM(A.销售额), A.地区,
  AVG(JSON_VALUE(B.BC_json,
    '$.buykpi')) AS 购买力指数
FROM CY销售数据_ALL A
LEFT JOIN TBL_BC_JSON B
ON A.地区= JSON_VALUE(B.BC_json,'$.dis')
GROUP BY A.地区;
```

使用Graph Analytics查找客户购买的共同商品

```
SELECT c1, e, p, e1, c2
FROM MATCH
   (c1)-[e]->(p)<-[e1]-(c2)
    on CUST_BUY
WHERE  c1.cust_id= 1246813
AND    C2.cust_id= 1002487
LIMIT 100;
```

使用Text Index和Search查找商品评论JSON中“延误”相关

```
SLECT ID, custid, jtext
FROM tbl_order_comments t
WHERE json_textcontains(t.jtext, '$.comments',
          '延误')
ORDER BY ID;
```

使用内置的Spatial功能查找某地铁站最近的10家店铺

```
SELECT A.ID, A.NAME, A.Long,
   A.Lat, sdo_nn_distance(1) as
   distance
FROM tbl_store A, tbl_station B
WHERE B.ID = 736 and
  SDO_NN(A.shape, B.shape,
'sdo_num_res=10',1) = 'TRUE'
ORDER BY distance;
```

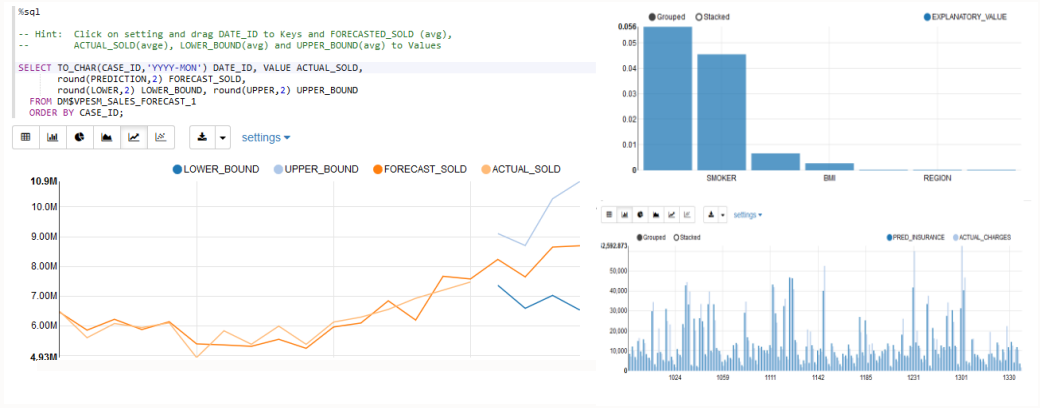
在Analytics View中动态浏览各层级汇总信息

```
SELECT
 sh_time_hier.member_name AS 时间,
 sh_time_hier.level_name as 时间层,
 sh_product_hier.member_name AS 产品,
 sh_product_hier.level_name as 产品层,
 AMOUNT_SOLD as 销售额
FROM  sh_sales_av HIERARCHIES
   (sh_time_hier, sh_product_hier)
```

使用创建的custbuy_mode预测客户购买商品的可能性

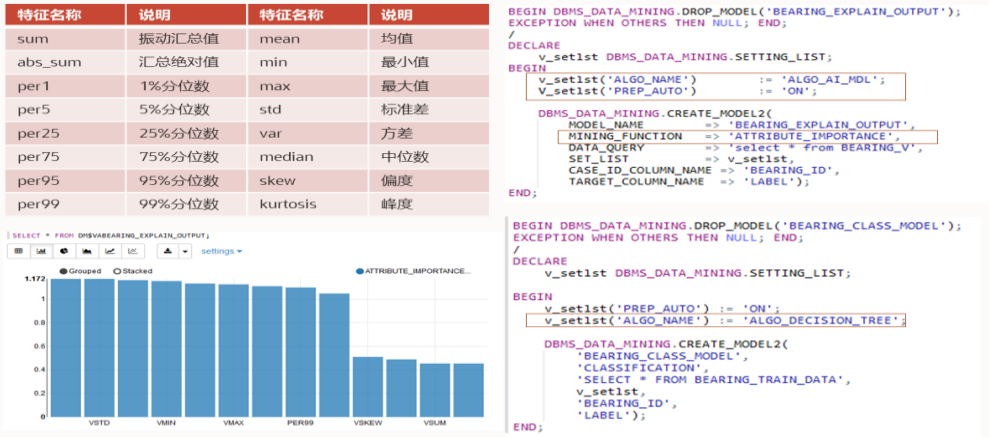
```
SELECT ROUND(PREDICTION_PROBABILITY(
   custbuy_model, 'P2' USING 42 as age,
   1 as car_type,
   2 AS cust_level,
   '男' AS gender,
   …… ),3) PROBABILITY_BUY
FROM DUAL;
```

Oracle数据库内机器学习和属性图分析能力

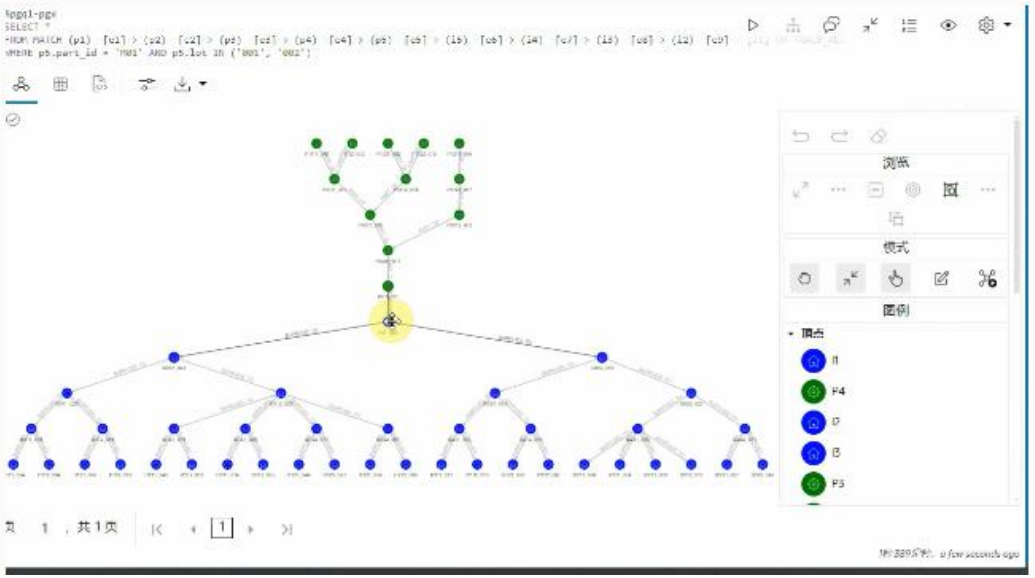


原材料价格和用量辅助决策

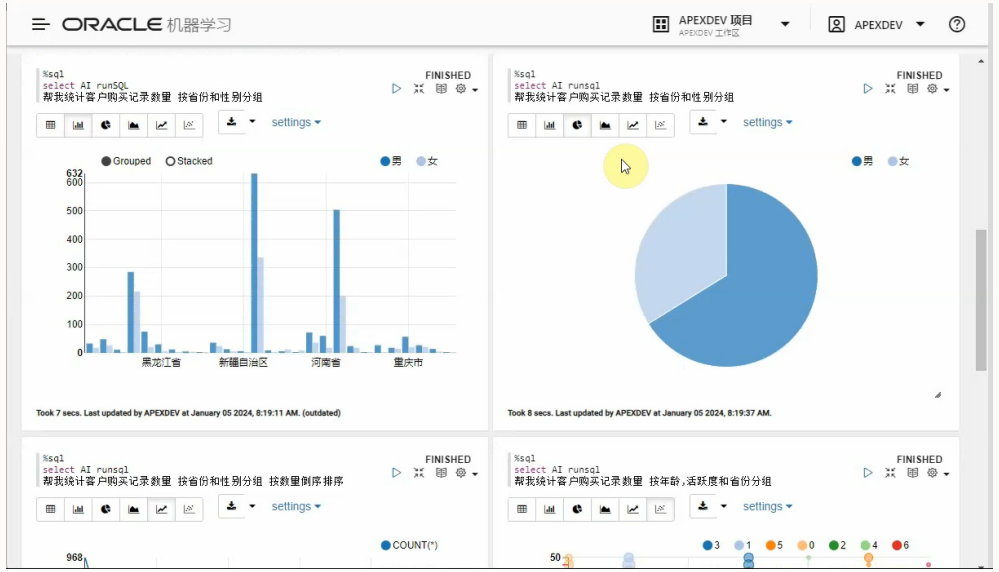
特征名称	说明	特征名称	说明
sum	振动汇总值	mean	均值
abs_sum	汇总绝对值	min	最小值
per1	1%分位数	max	最大值
per5	5%分位数	std	标准差
per25	25%分位数	var	方差
per75	75%分位数	median	中位数
per95	95%分位数	skew	偏度
per99	99%分位数	kurtosis	峰度



设备零件失效影响原因相关性探查



实体关系探查（BOM变更，产品关联）



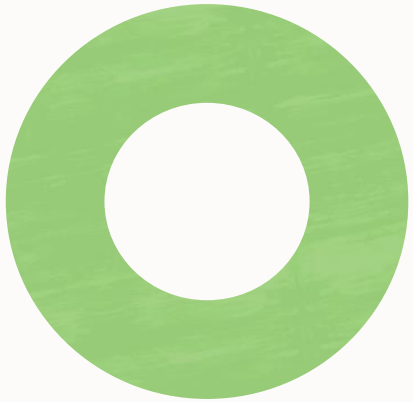
Select AI/检索增强生成

总结： Oracle AI不同使用场景的产品技术及特点

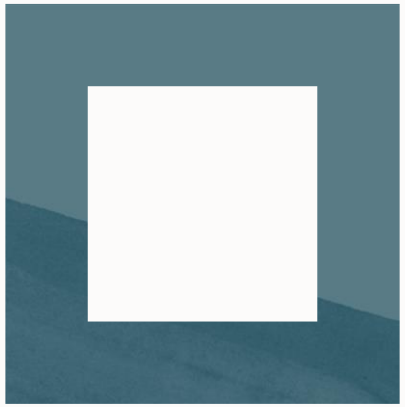
		Oracle提供的服务和产品技术	特点
9	机器学习和数据平台	Oracle Database，ML4SQL/R/Python Exadata / Exadata Cloud@Customer	开箱即用经典算法，按需部署，有制造行业实际案例检验和最佳实践
8	检索增强生成（RAG）/知识库/向量数据库	Oracle Database, OCI Heatwave	多模数据库，结合客户业务和其他数据，安全，易用，适用广泛
7	生成式AI和SQL结合	OCI ADB	开箱即用，结合客户表结构
6	应用和场景	Oracle SaaS(内置AI能力)	开箱即用，结合客户数据
5	应用中使用AI服务能力	OCI AI Service + 客户数据 (视频+语音+文本+机器人)	开箱即用，API或界面，结合Data Labeling对数据标注和训练
4	自定义模型	OCI 生成式AI+客户数据	可针对客户数据微调，数据隔离
3	开箱即用的Gen. AI服务	OCI 生成式AI服务	开箱即用，企业级，可独立服务器， 按需计费
2	模型	*Cohere，开源，* OCI Data Science开发，训练，部署	协同开发平台，简化开发训练和部署过程
1	算力	OCI SUPPER Cluster/GPU/CPU	性价比，高速联通网络，资源

Oracle’s AI is…

专注于企业



完整



隐私和安全

数据库和云讲座群

20-23

甲骨文云技术公众号

技术专家1V1深入交流


ORACLE